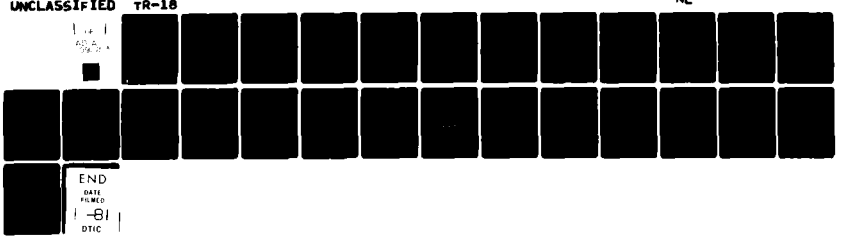
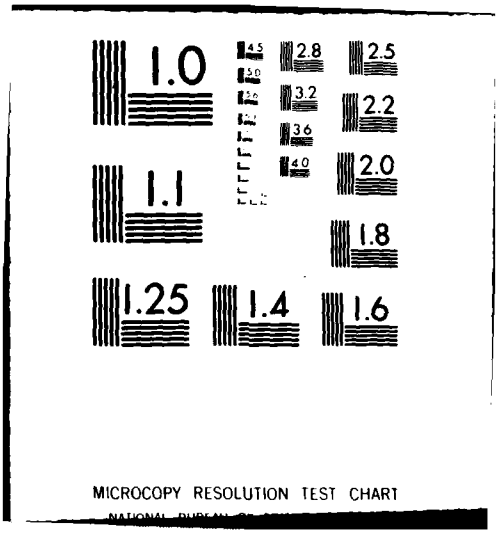
AD-A092 213 WASHINGTON UNIV SEATTLE LAB FOR CHEMOMETRICS F/G 7/4
IFVA: A COMBINED LINEAR AND NONLINEAR FACTOR ANALYSIS PROGRAM P--ETC(U)
NOV 80 C JOCHUM, B R KOWALSKI N00014-75-C-0536
UNCLASSIFIED TR-18 NL



END
DATE FILMED
1-81
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TR-18	2. GOVT ACCESSION NO.: AD-A092213
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): UFVA, A Combined Linear and Nonlinear Factor Analysis Program Package for Chemical Data Evaluation.	5. TYPE OF REPORT & PERIOD COVERED: Technical Report - Interim 8/80 - 11/80
	6. PERFORMING ORG. REPORT NUMBER: 15
7. AUTHOR(s): Clemens Jochum, Bruce R. Kowalski	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0536
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 051-565
11. CONTROLLING OFFICE NAME AND ADDRESS: Materials Sciences Division, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: November 1980
	13. NUMBER OF PAGES: 26
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Interim rept. Aug-Nov 80	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Prepared for publication in Analytica Chimica Acta

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

underlying variable factor analysis
factor loading matrix; factor weight matrix

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

UFVA, an underlying variable factor analysis program is described. The theories of principal component analysis and nonlinear least squares projection techniques are outlined and compared. Several applications from various chemical fields are presented which show that a complete analysis of the underlying structure and dimensionality of a chemical data set should always include these nonlinear projection techniques.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

OFFICE OF NAVAL RESEARCH

Contract N00014-75-C-0536

Task No. NR 051-565

TECHNICAL REPORT NO. 18

UFVA, A Combined Linear and Nonlinear Factor Analysis Program Package for Chemical Data Evaluation

by

Clemens Jochum and Bruce R. Kowalski

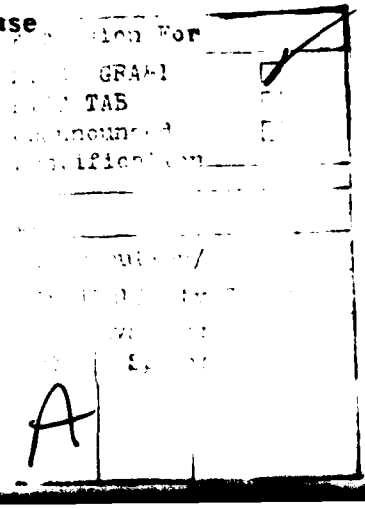
Prepared for Publication

in

Analytica Chimica Acta

Laboratory for Chemometrics
Department of Chemistry BG-10
University of Washington
Seattle, Washington 98195

November 1980



SUMMARY

UVFA, an underlying variable factor analysis program is described. The theories of principal component analysis and nonlinear least squares projection techniques are outlined and compared. Several applications from various chemical fields are presented which show that a complete analysis of the underlying structure and dimensionality of a chemical data set should always include these nonlinear projection techniques.

INTRODUCTION

Multivariate statistics, originally developed for applications in social sciences, have been more and more applied to chemical data evaluation. In fact, the statistical treatment of chemical data became a whole new branch of analytical chemistry, called Chemometrics [1].

One of the most powerful methods in chemometrics which has been applied as a "stand-alone" method as well as in combination with other methods is principal component factor analysis [2] PCA. Applications range from data reduction problems, interpretation of the underlying structure of a data set to a preliminary treatment of the data bases for a path modelling analysis [3]. PCA has been applied to e.g. mass spectral and environmental data, NMR and chromatography data [4].

PCA assumes a linear relation among the variables. In nature, however, most relations between physical parameters or variables are nonlinear. To overcome this setback of linear factor analysis, algorithms such as nonlinear least squares, multidimensional scaling [5] and parametric mapping [6] for the analysis of the underlying nonlinear structure of a data base have been developed. So far, there have been no applications published of these nonlinear methods to chemical data analysis. In the next section the theory of the different linear and nonlinear methods is explained.

In the following an interactive program package is described which includes not only principal component factor analysis and rotational methods, but also nonlinear least squares projection techniques such as multidimensional scaling, nonlinear and parametric mapping and graphical output routines. The algorithms and the program are demonstrated on two chemical data sets.

THEORY

The underlying relation among n measurements (e.g. melting point, dipole moment, etc.) of a data matrix $Z = (Z_{ij})\ \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}$ consisting of m samples is to be analyzed.

To give the measurements equal weight, they are usually scaled to unit variance and zero mean.

In a three variable data set (n = 3) the measurement vectors can be represented graphically in a three-dimensional space (Fig. 1).

Figure 1

Factor analysis determines the dimensionality of the hyperspace necessary to represent the data. The first factor λ_1 is represented by the longest axis of the hyperspace containing the data, i.e. it represents the largest amount of variance in one dimension. The second vector λ_2 is represented by the second longest axis orthogonal to the first one and so on. To obtain the r factors necessary to represent most of the total variance of the data set, the data matrix Z is decomposed into a factor weight matrix (factor loading matrix) $A = (a_{ij})\ \begin{matrix} i = 1,\ldots,n \\ j = 1,\ldots,r \end{matrix}$ and a factor score matrix $P = (p_{ij})\ \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,r \end{matrix}$ ($r \leq n$):

$$Z = P \cdot A^T$$

The columns of A are determined by calculating the eigenvectors of the data covariance matrix C

$$C = Z^T Z.$$

The entries a_{ij} of the factor loading matrix A can be considered as the multiple correlation coefficients of the variable i with the factor j.

The factor score matrix represents the data in terms of factor coordinates and is calculated according to

$$P = Z\,A.$$

This transformation is known as the Karhunen-Loeve expansion [7].

As mentioned above, the relation between physical parameter variables is not always linear. It is, for example, possible that the data lie along a curved line or surface (Fig. 2).

Figure 2

Linear principal component factor analysis would still come up with three factors since the variance for all possible three dimensional orthogonal coordinate systems is greater than zero in any coordinate direction. On the other hand, there are obviously only two underlying nonlinear independent variables.

To solve this problem, nonlinear least squares projection methods have been developed [5,6].

The distances (d_{ij}) $i,j = 1,\ldots,m$ between all m data vectors of Z are calculated. The data points are then arranged in an r-dimensional space $(r<n)$ in a way that the stress S

$$S = \sqrt{\frac{\sum_{i,j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

is minimal [5]. The $\hat{d}_{ij}$ denote the recalculated distances of the data points in the lower r-dimensional space. The stress S thus represents a measure of the goodness of fit of the data vectors projected in the r-dimensional space compared with their configuration in the original n dimensional space.

The different projection techniques differ mainly by a different measure for the goodness of fit. To demonstrate the different applications for PCA, multidimensional scaling [5] and parametric mapping [6], their optimum theoretical results on three different two dimensional data sets (I, II, III) are shown (Fig. 3)

Figure 3

The parametric mapping algorithm is able to determine a ring shaped one dimensional structure of the data since it considers only local environments of data points. Since this method does not look at the global fit of all data points it, however, sometimes ends up with a too small dimensionality.

Although there exist programs for these nonlinear least squares methods, they are not set up for chemical data bases. They are not input compatible with each other and they work only as batch programs. Since these programs only include either multidimensional scaling or parametric mapping and no linear factor analysis program, there was a definite need for a combined package. Such a combined underlying variable factor analysis program is described in the next section.

THE PROGRAM UVFA

The underlying variable factor analysis program UVFA [9] consists of a driver routine, a set of utility routines and 21 major subroutines which perform the actual data analysis. Since the data are stored on disk files, only the driver routine and the utility routines have to stay in core during the whole run. The 21 major subroutines can be loaded one at a time. Thus the program usually needs less than 60_8 K words of core to run although it consists of more than 10,000 statements.

The input, output and internal binary files are fully compatible with our pattern recognition program ARTHUR [8].

UVFA can be run interactively or in batch mode and has graphical output routines for Tektronix 4010/4014 terminals, Calcomp plotter or line printer. Figure 4 shows the general setup of the program.

Figure 4

PRICO does a principal component analysis with or without communality iteration [2]. MULSCA and PARAMA are the nonlinear least squares projection routines for multidimensional scaling and parametric mapping. The underlying linear and nonlinear factors can be plotted with the routines PRIPLO (line printer plot), CALPLO (Calcomp plot) or TEKPLO (Tektromix graphics terminal plot). For additional error analysis the linear factors can be backtransformed by calling KATRAN (Karhunen-Loeve-Transformation) and BACKTR. The program also assists with the interpretation of the factors by calling ANALYS (ordering of the factors and loadings and performing various tests for finding the intrinsic dimensionality), HIER (performing a hierarchical cluster analysis), ROTOR and ROTSUB for performing various kinds of rotations.

There exist versions for 60 bit CDC computers and 32 bit DEC VAX computers. The VAX version should be well compatible with other DEC and IBM computers. The whole program is written in FORTRAN.

APPLICATIONS

Among the various applications, three are discussed in more detail: A mass spectral data set, a constitutional similarity set of chemical compounds and a data set of physical parameters of biologically interesting compounds.

The first data set consists of the mass spectra of 11 mono and sesquiterpenes [10]. These are Isoprene (1), Myrcene (2), p-Cymene (3), β-Pinene (4), Camphene (5), Limonene (6), α-Cedrene (7), Caryophyllene (8), β-Selinene (9), Santene (10), δ-Cadinene (11). Figure 5 shows the plot of the loadings of the first two vectors of the factor weight matrix.

Figure 5

These two factors encounter 97% of the total variance. We see two clusters of compounds; only compound 8 seems to lie somewhat in between. It turns out that one cluster consists of the monoterpenes and Isoprene; the second is of the sesquiterpenes. Compound 8 (Caryophyllene) should therefore belong to the second cluster (see below). Since the first factor encounters already 94% of the total variance there is clearly one main factor, i.e. there is one main underlying fragmentation pattern.

The nonlinear multidimensional scaling configuration of our data in two dimensions shows the separation of the two clusters very clearly (Fig. 6).

Figure 6

The very similar fragmentation pattern of Isoprene and the monoterpenes is reflected by their close neighborhood within the cluster. The one dimensional multidimensional scaling of the mass spectra (Fig. 7) corroborates that there is mainly one underlying fragmentation pattern: The stress of the one dimensional projection is almost as low as for two dimensions [11] (0.0031 and 0.008 respectively) and thus the intrinsic dimensionality is most likely one.

Figure 7

In our second example, the data base consists of a distance matrix $D = (d_{ij})$ $i,j = 1,\ldots,13$ of another set of 13 terpene components. These are Isoprene, four monoterpenes (Myrcene, Menthol, Camphene, Umbellulone), four sesquiterpenes (Bisabolene, α-Cadinol, Eudesmol, Partheniol), three Diterpenes (Dextropimaric Acid, Phyllocladene, Roylleanone) and one Triterpene (β-Amyrin) [12]. The distance measure d_{ij} is the minimum chemical distance [13] between the compounds i and j. It indicates the constitutional similarity between two compounds. To perform a principal component analysis, a covariance matrix C is generated from the distance matrix [14]:

$$d_{ij} = 2\ (1 - C_{ij})^{1/2}$$

We again get two linear factors (85.8% and 13.1% partial variances) and a plot of their loadings (Fig. 8) shows that there are no particular clusters of compounds.

Figure 8

For further interpretation we look at the two dimensional nonlinear projection (Fig. 9).

Figure 9

The compounds are now clustered according to whether they are Mono, Sesqui, Di- or Triterpenes. Again the stress for a one dimensional projection is almost as low as for two dimensions (0.0087 and 0.00859, respectively) which suggests that all these compounds are built by one major structural element, the isoprene unit. This is indicated by the ordering of the compounds along the one dimensional projection axis (not shown) according to their number of isoprene units.

The third example corroborates some results of a linear factor analysis (principal component analysis) performed by R. D. Cramer on a data set of 10

physical parameters of 44 organic compounds [15]. Cramer obtains two linear factors with 75.5% and 21% partial variance. A nonlinear projection of the compounds in a two dimensional space shows no distinct clusters of the compounds (Fig. 10).

Figure 10

Since the data points are not lying along a line but they are spread almost equally in both directions, the intrinsic dimensionality seems to be two. An attempt to project the data in a one dimensional space results in a very interesting pattern of a plot of the calculated distances $\hat{d}_{ij}$ versus the original d_{ij} (see above) of the configuration (Fig. 11).

Figure 11

Although most of the distance points lie close to the diagonal which indicates that most of the compounds can be fairly well fit in one dimension, some of them lie almost along an axis perpendicular to the diagonal. This most likely indicates that there is one major and one minor nonlinear factor in the two dimensional space. This is, however, subject to further investigation.

CONCLUSION

The examples show that the combination of principal component analysis with nonlinear least squares projection techniques is a very powerful tool for the determination of the intrinsic dimensionality and the interpretation of the factors. Our underlying variables factor analysis program UVFA [9] provides a convenient way for a complete factor and nonlinear projection analysis of any data set.

CREDIT

This work was supported in part by the Office of Naval Research. The work was also supported in part by the German Academic Exchange Service.

REFERENCES

1. B. R. Kowalski (Ed.), Chemometrics: Theory and Application, Vol. 52, ACS Symposium Series, 1977.
2. Since communality estimate and iteration is of minor importance for chemical applications, we do not distinguish between principal component and linear factor analysis. For more detailed information on this subject refer, for example, to K. Überla, Faktorenanalyse, Springer-Verlag, Berlin, 1977.

3a. R. W. Gerlach, B. R. Kowalski, and H. Wold, Anal. Chim. Acta, 112 (1977) 417.

b. R. W. Gerlach, Multivariate Methods in Chemistry, Ph.D. Thesis, University of Washington, Seattle, 1980.

4. E. R. Malinowski, D. G. Howery, Factor Analysis in Chemistry, Wiley-Interscience, New York, 1980.

5a. J. B. Kruskal, Psychometrika, 29 (1964) 1.

b. J. B. Kruskal, Psychometrika, 29 (1964) 115.

6. R. N. Shepard, J. D. Carroll, in P. R. Krishnaiah (Ed.), International symposium of multivariate analysis, Academic Press, New York, 1966, p. 561.
7. H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Wiley-Interscience, New York, 1972.
8. D. L. Duewer, J. R. Koskinen, B. R. Kowalski, Documentation for ARTHUR, Version 1-8-75, Chemometrics Society Report No. 2, Seattle, 1975.
9. A well documented version of the program is available from Infometrix, Inc., P. O. Box 25808, Seattle, Washington, 98125, USA.
10. E. Stenhagen, S. Abrahamsson, F. W. McLafferty (Eds.), Atlas of Mass Spectral Data, J. Wiley and Sons, New York, 1969.
11. Kruskal [5b] considers a stress of 0.1 to 0.05 as "satisfactory" and below 0.05 as "impressive". He, however, cautions that a low stress is only a necessary criteria and that a meaningful interpretation of the configuration is most important.

12. J. B. Hendrickson, The Molecules of Nature, W. A. Benjamin, Inc., New York, 1965.

13. C. Jochum, J. Gasteiger, I. Ugi, Angew. Chemie Int. Ed., 19 (1980) 495.

14. A. P. M. Coxon, C. L. Jones, in C. A. O'Muircheartaigh, C. Payne (Eds.), Exploring Data Structures, Vol. 1, J. Wiley, London, 1977, p. 174.

15a. R. D. Cramer III, J. Am. Chem. Soc., 102 (1980) 1837.

b. R. D. Cramer III, J. Am. Chem. Soc., 102 (1980) 1849.

FIGURE CAPTIONS

Figure 1: Three dimensional representation of measurement vectors.

Figure 2: Three dimensional data points lying on a nonlinear surface.

Figure 3: Comparison of Principal Component, Multidimensional Scaling and Parametric Mapping Analysis.

Figure 4: General Schematic of the UVFA program.

Figure 5: Plot of the loadings of the first two factors of the terpene mass spectral data.

Figure 6: Two dimensional nonlinear projection of terpene mass spectral data.

Figure 7: One dimensional projection of terpene mass spectral data.

Figure 8: Factor loadings of minimum chemical distance data of 13 terpenes.

Figure 9: Two dimensional projection of the minimum chemical distance terpene data.

Figure 10: Two dimensional nonlinear projection of 10 physical parameters/ 44 compound data.

Figure 11: Original versus calculated distances of the one dimensional optimum configuration of the 10 physical parameters/44 compound data.

Figure 1

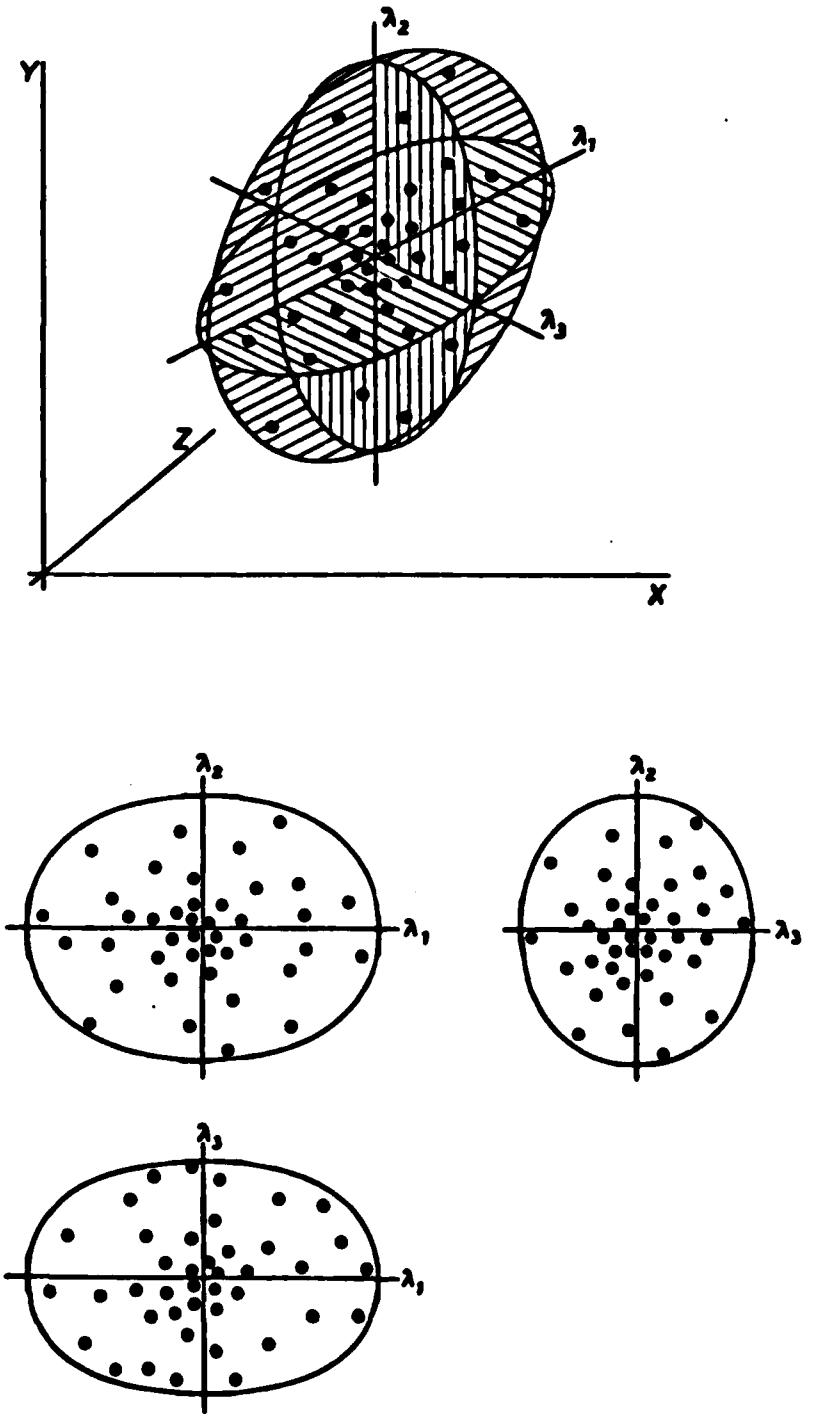


Figure 2

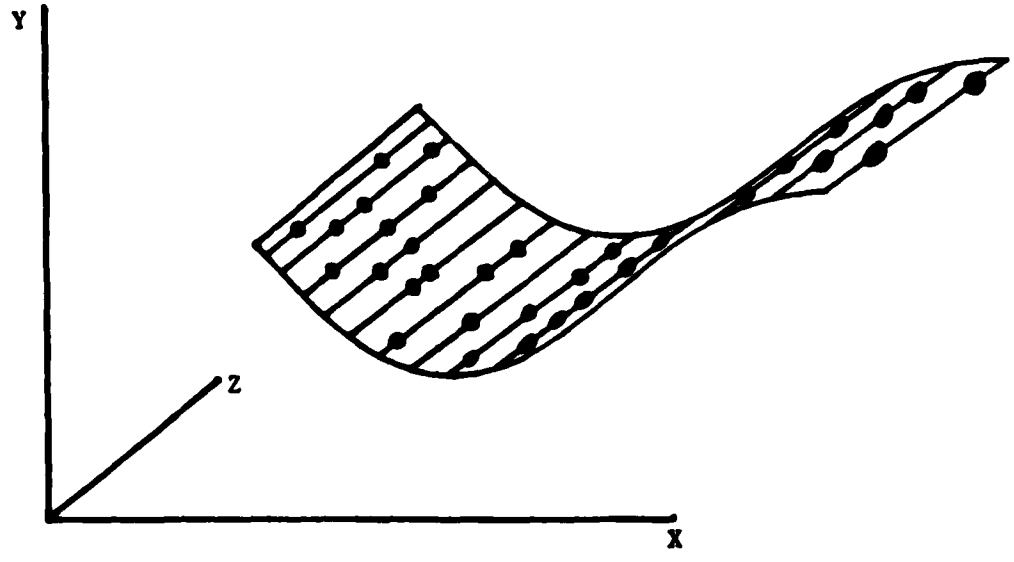
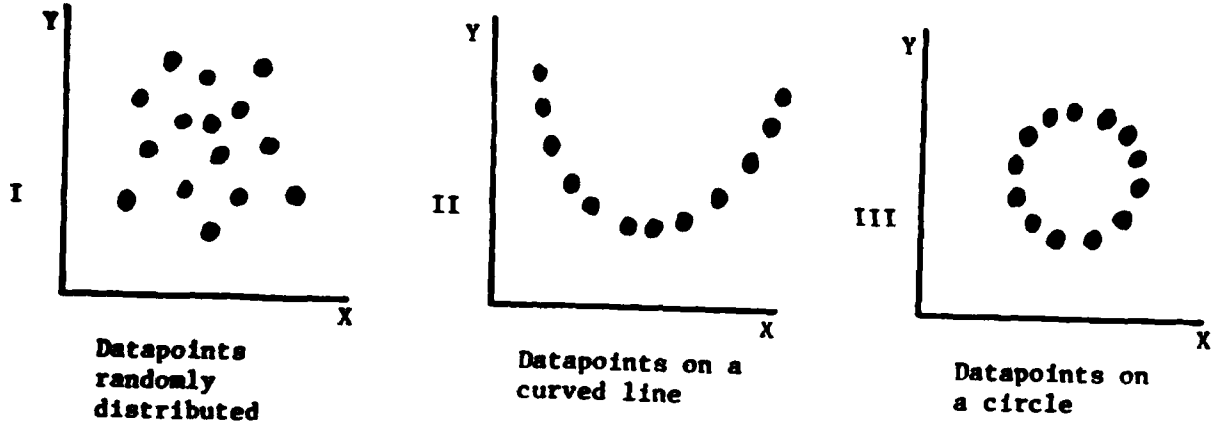


Figure 3



Dataset	Dimensionality found by:		
	Principal Component Analysis	Multidimensional Scaling	Parametric Mapping
I	2	2	2
II	2	1	1
III	2	2	1

Figure 4

General Underlying Variable Factor Analysis Schematic

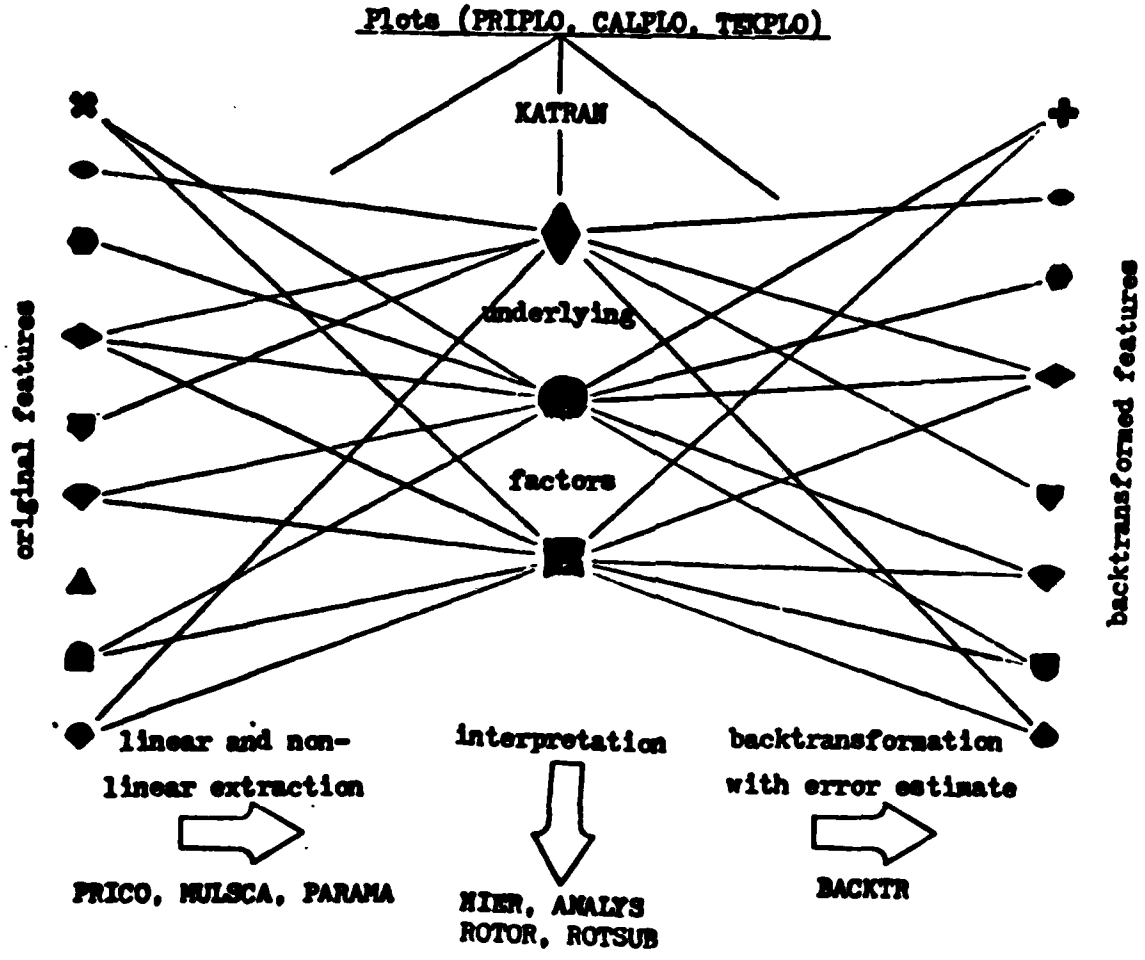


Figure 5

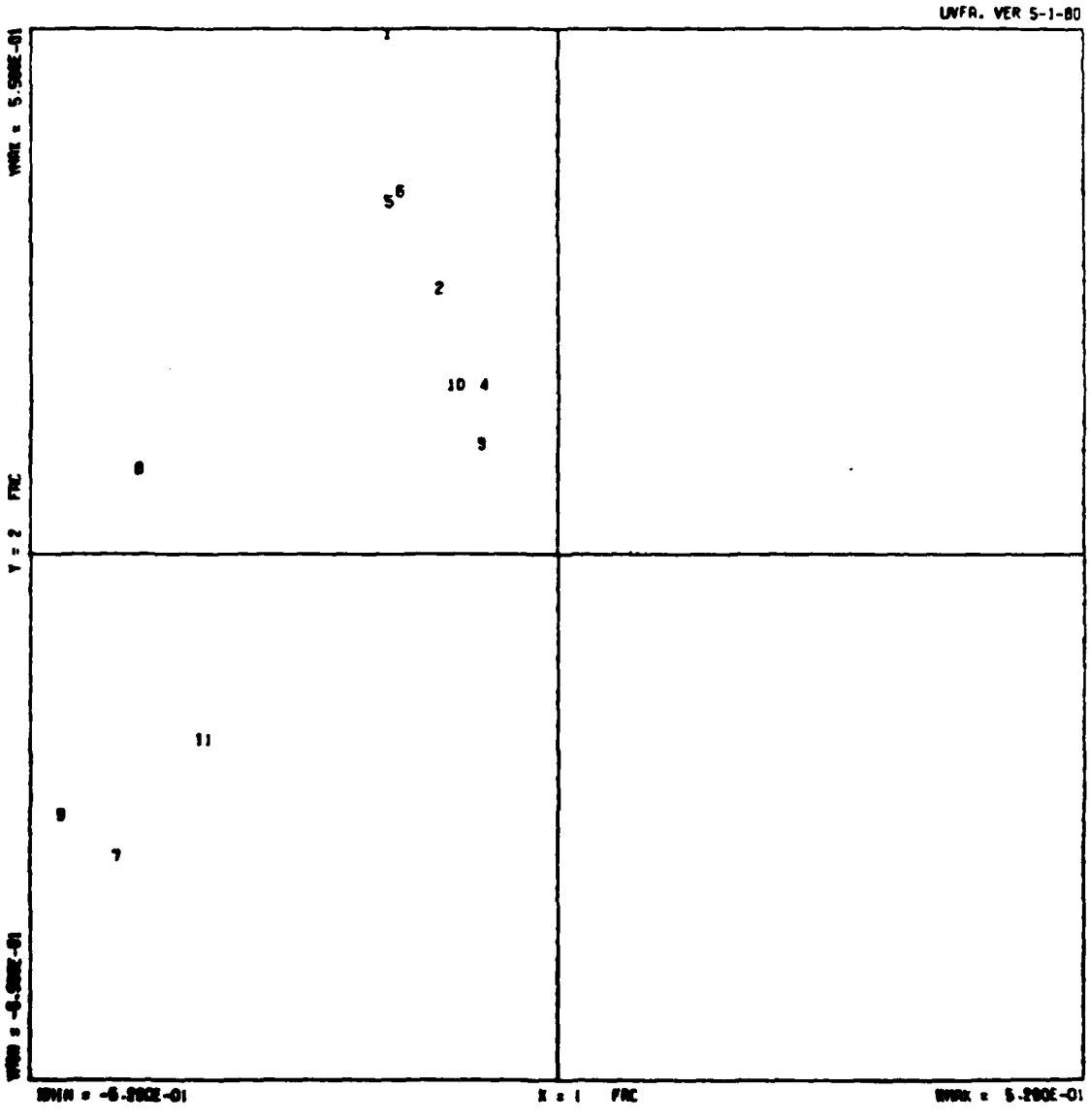


Figure 6

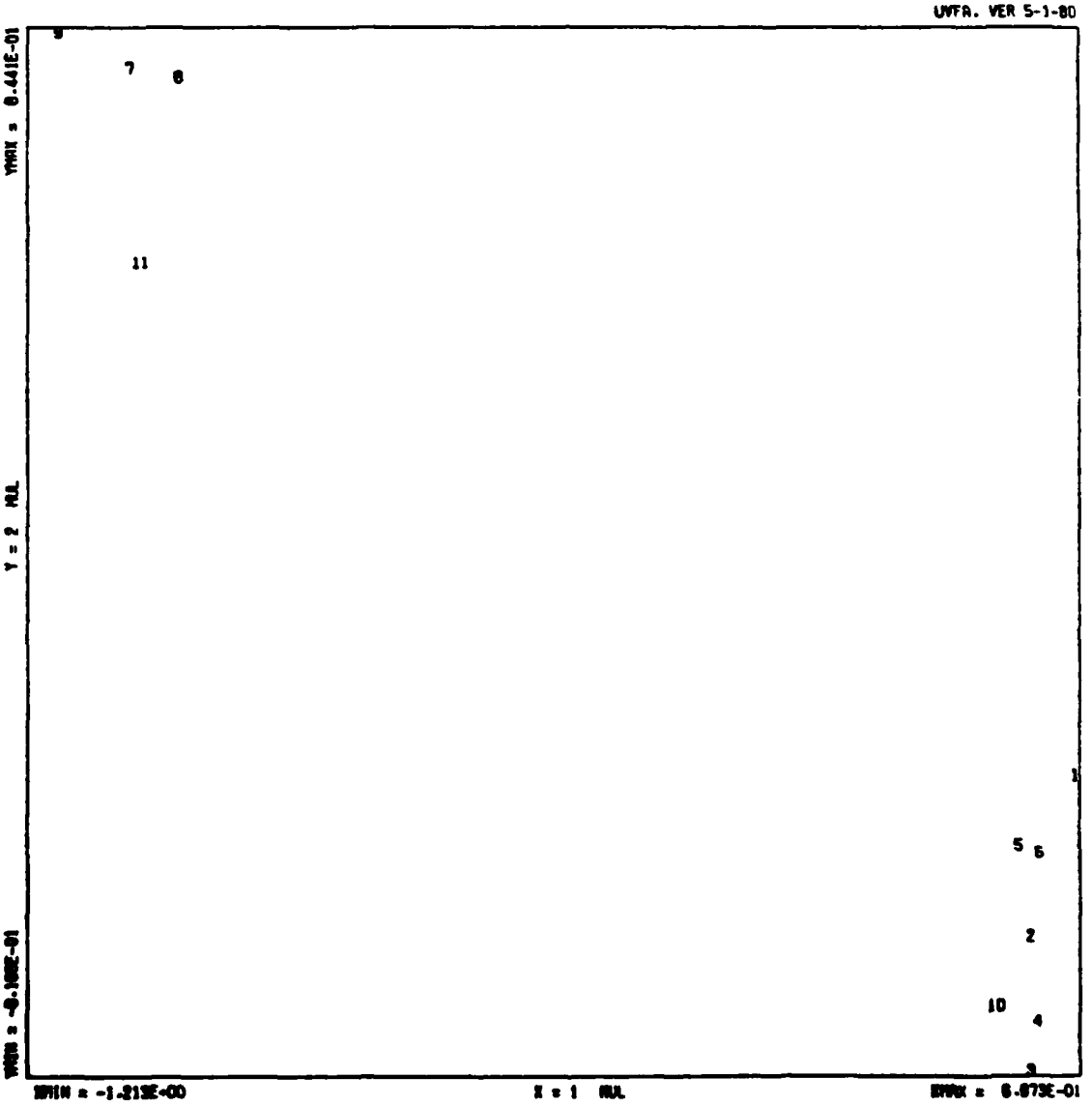

UVFA. VER 5-1-80
YMAX = 0.441E-01
Y = 2 NUL
YMIN = -0.308E-01
XMIN = -1.215E+00
X = 1 NUL
XMAX = 0.873E-01
9
7
8
11
1
5
6
2
10
4
3

Figure 7

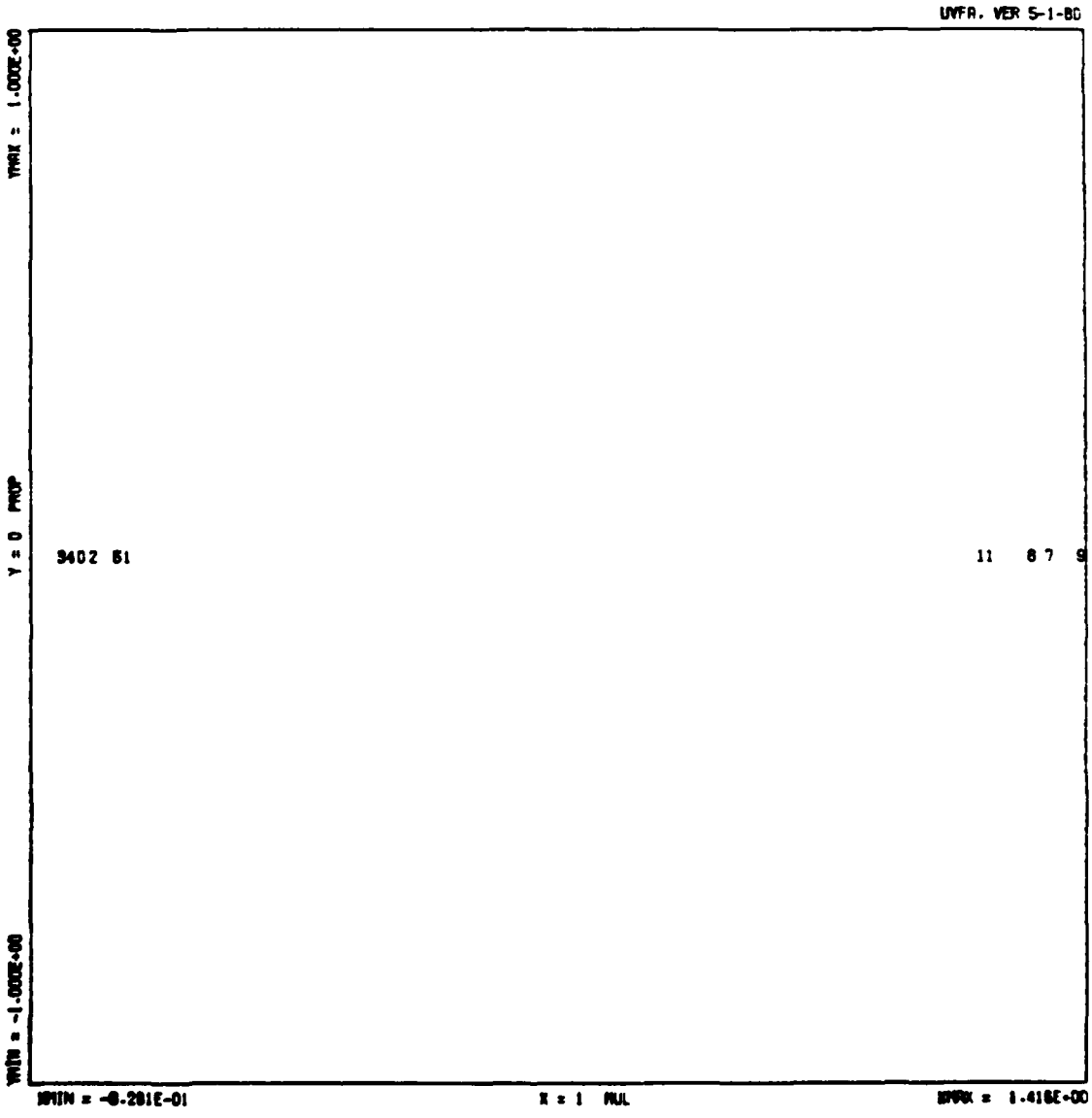

UVFA, VER 5-1-80
YMAX = 1.000E+00
Y = 0 PROP
YMIN = -1.000E+00
XMIN = -0.281E-01
X = 1 MUL
XMAX = 1.416E+00

Figure 8

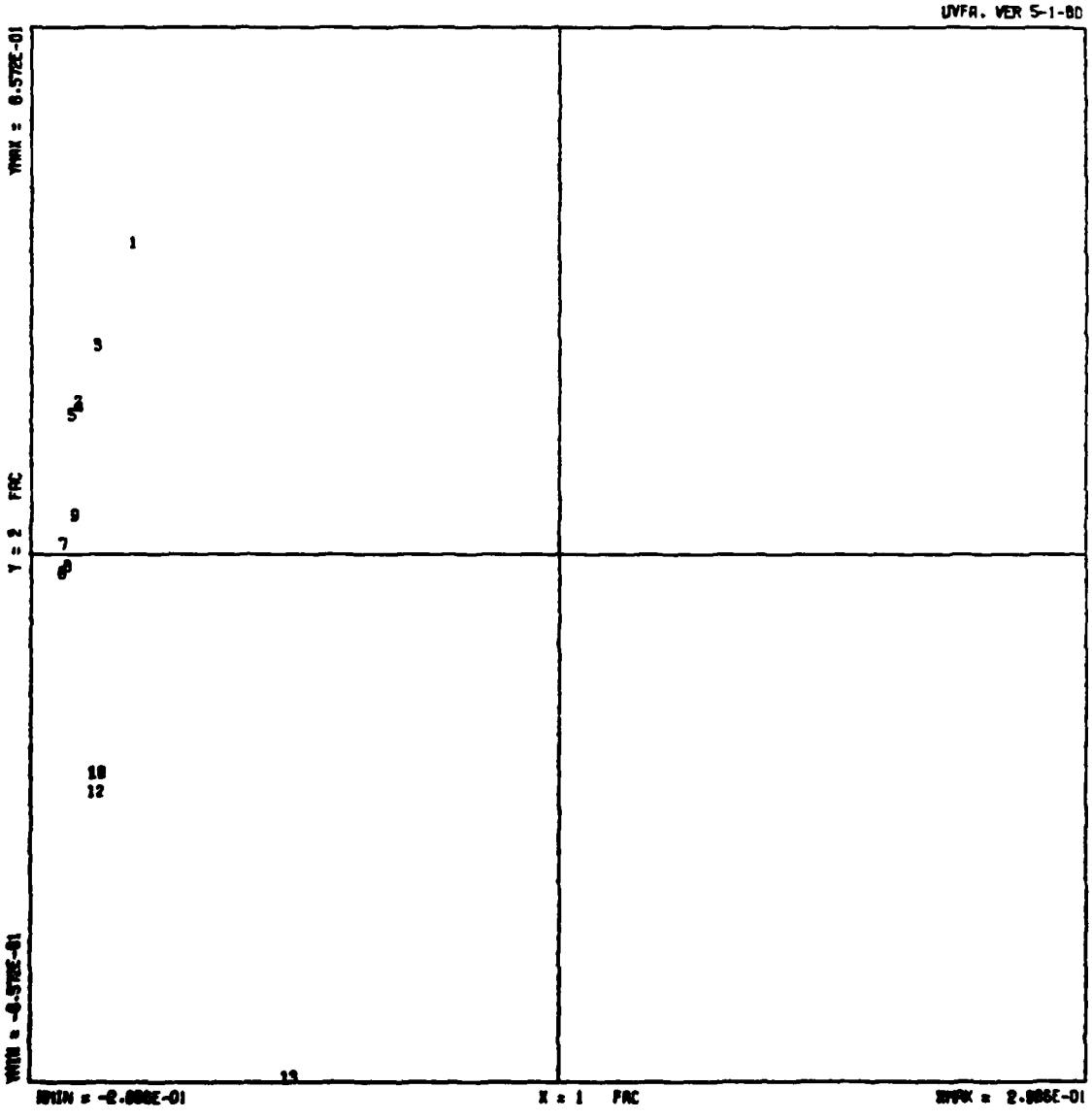

UVFA. VER 5-1-80
YMAX = 8.572E-01
Y = 2 FAC
YMIN = -8.572E-01
XMIN = -2.888E-01
X = 1 FAC
XMAX = 2.888E-01
1
3
2
5
9
7
6
10
18
12
13

Figure 9

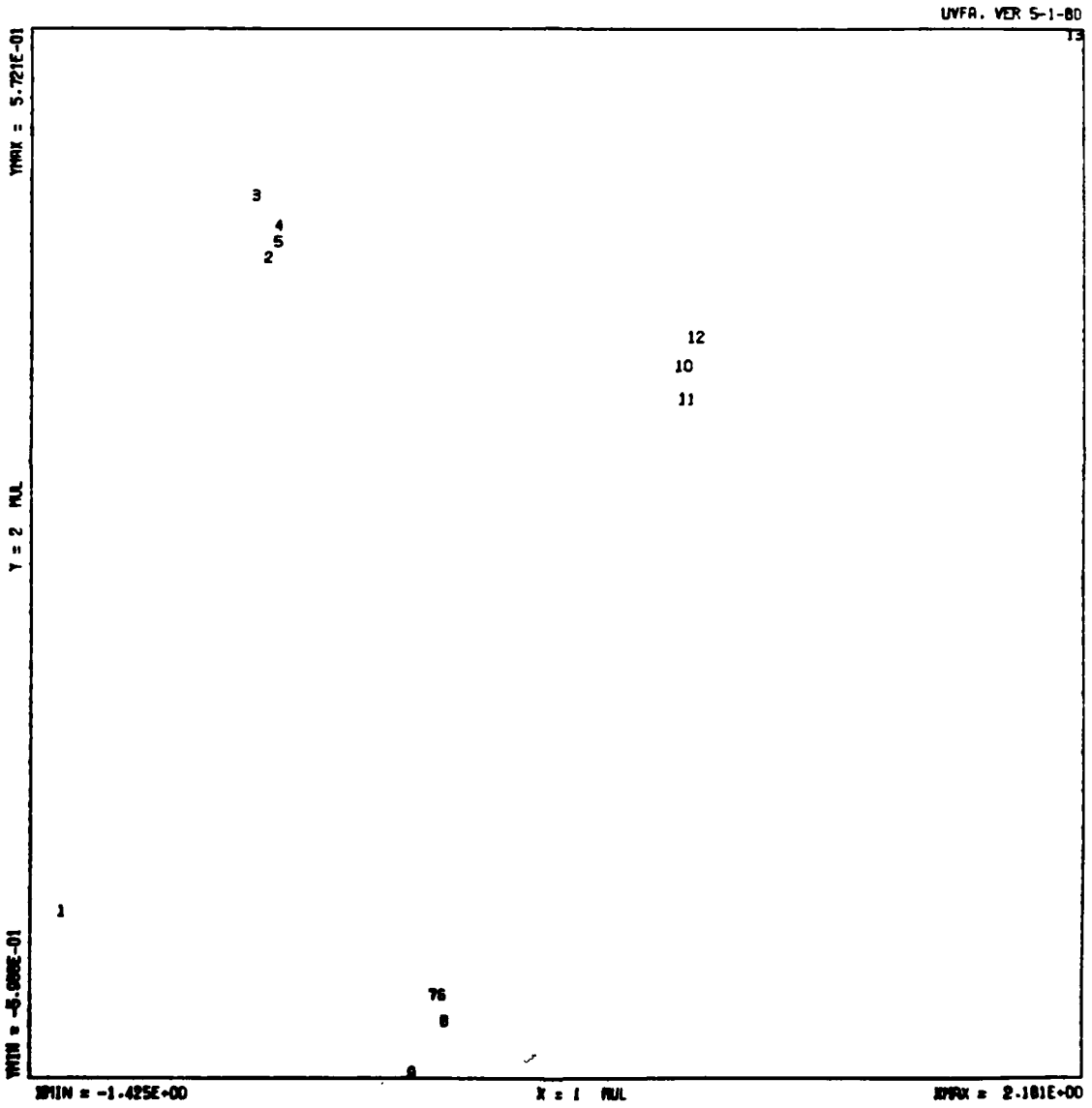

UVFA. VER 5-1-80
YMAX = 5.721E-01
YMIN = -5.000E-01
Y = 2 MUL
XMIN = -1.425E+00
X = 1 MUL
XMAX = 2.181E+00
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 10

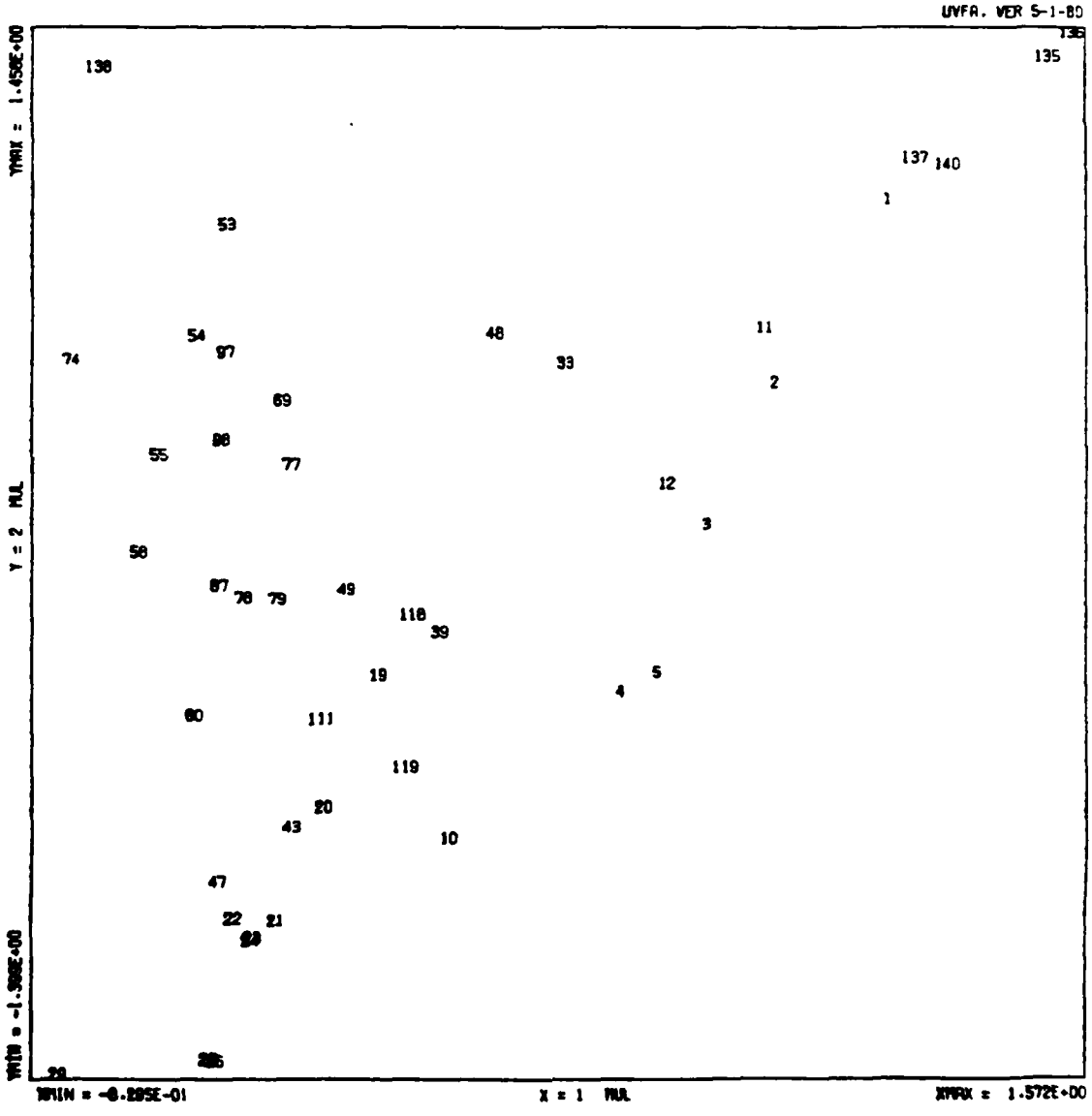
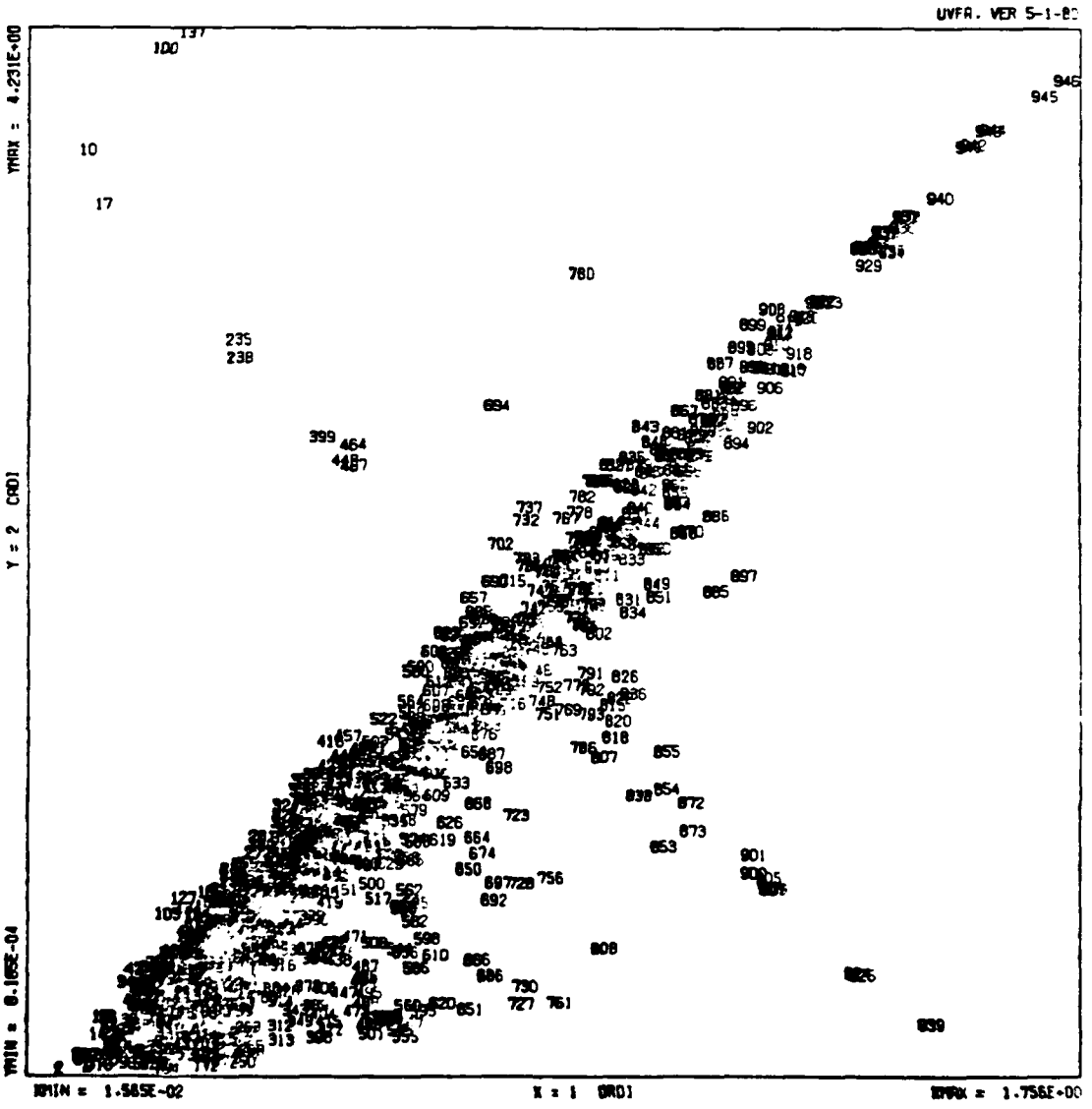

UVFA. VER 5-1-80
YMAX = 1.450E+00
YMIN = -1.300E+00
Y = 2 PUL
XMIN = -8.285E-01
X = 1 PUL
XMAX = 1.572E+00
136
135
138
137
140
1
53
54
48
11
74
97
93
2
69
88
55
77
12
3
58
87
78
79
49
118
39
19
5
4
60
111
119
20
43
10
47
22
21
28
285
29

Figure 11


UVFA. VER 5-1-8
YMAX = 4.231E+00
Y = 2 ORD1
YMIN = 8.165E-04
XMIN = 1.585E-02
X = 1 ORD1
XMAX = 1.756E+00


DAT
ILM